学校编码：10384　　分类号_____ 密级 _____

学号：X2011230668　　UDC _____

厦门大学

工 程 硕 士 学 位 论 文

汉—哈萨克双语电子词典的设计与实现

Design and Implementation of Chinese-Kazakh Diglossia Electronic Dictionary

古丽孜亚·阿布都吉力

指导教师姓名：龙 飞 副 教 授

专 业 名 称：软 件 工 程

论文提交日期：2013 年 10 月

论文答辩日期：2013 年 11 月

学位授予日期：2013 年 月

指 导 教 师：_____

答辩委员会主席：_____

2013 年 10 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

本文主要论述了汉哈萨克双语电子词典构建和相关技术的研究。汉哈萨克双语电子词典的构建是为 Windows 平台下实现汉语和哈萨克语词汇间相互查询功能，同时实现 Windows 平台下为其他文字处理系统提供辅助功能的屏幕取词技术，利用该技术可以实现鼠标即点下的汉语词汇和哈萨克语单词的解释。该电子词典提供词汇查询、词库维护、屏幕取词、即时互译等功能。

利用微软公司 Office Access 来构建汉哈萨克双语电子词典的词库。由于词典的编撰需要耗费大量的人力和物力，所以在国内，内容丰富和全面的汉哈萨克双语词典种类不多，同时内容更新较慢，因此本次在最新一次编撰的《汉哈辞典》的基础上建立了汉哈萨克和哈萨克汉双语词库。同时将该电子词典的用户定义为大学和中小学的学生群体。

词典作为语言学习的工具，最主要的功能就是查询，所以汉哈萨克双语电子词典可以进行两种方式的查询，一种是通过键盘输入词汇的形式进行汉语和哈萨克语的词义查找，另一种形式是通过屏幕取词技术，让计算机程序自动获取所指词汇，通过判断是汉语还是哈萨克语来进行查找词义并显示。

在汉哈萨克双语电子词典中，重点讨论了关于哈萨克语在自然语言处理领域的几个重要问题，一个是 Unicode 代码显示，一个是分词技术，另一个是词干提取技术。同时也对汉语的自动分词技术进行了讨论。

关键字：汉哈萨克双语；电子词典；词干提取

Abstract

This article discusses the “Chinese-Kazakh bilingual electronic dictionary” of the building and related technologies. " Chinese-Kazakh bilingual electronic dictionary ," the building is for the Windows platform to achieve mutual Chinese and Kazakh vocabulary query capabilities , while enabling the Windows platform for other word processing system provides accessibility Screen technology , the mouse can be achieved using this technique that point the Chinese vocabulary and Kazakh word explanation. This electronic dictionary provides vocabulary queries, thesaurus maintenance, capturing, instant translation and other functions.

Use Microsoft Office automation software suite Access to build “Chinese-Kazakh bilingual electronic dictionary” thesaurus. Since dictionary compilation takes a lot of manpower and material resources, so that domestic content-rich and comprehensive Chinese-Kazakh bilingual dictionary narrow range , while content updates slow, so this in the latest compilation of the " Dictionary of Chinese and Kazakh " established on the basis the Hanhasake and Kazakh Chinese bilingual thesaurus. This electronic dictionary while schools and universities and user-defined groups of students.

Dictionary as a language learning tool, the most important function is to query, so " Chinese-Kazakh bilingual electronic dictionary " queries can be performed in two ways, one is through the keyboard input in the form of words of Chinese and Kazakh find meaning, and the other forms by capturing technology that allows a computer program within the meaning of words automatically get through to judge Chinese or Kazakh to find meaning and displayed.

In the “Chinese-Kazakh bilingual electronic dictionary ", the Kazakh focused on the field of natural language processing in several important issues , one is the Unicode code shows , one word segmentation , the other is stemming technology . But also on the Chinese word segmentation techniques are discussed.

Key words: Chinese-Kazakh Diglossia; Electronic Dictionary; Stemming

目录

Contents

第一章　绪论

1.1 课题背景及意义

在自然语言处理领域中，电子词典一直是其最重要的基础，其为自然语言的其他处理工作提供了最基础的知识来源[1][2]。随着少数民族语言对信息化的要求越来越高，针对哈萨克语和其他语言间的机器翻译系统的开发也开始得到重视，作为最重要的对哈萨克语词汇的研究越来越被计算语言学所重视。对提高机器翻译的质量，电子词典的作用至关重要。在新疆对哈萨克语的自然语言处理才刚刚起步，方兴未艾，如何研究和实现一个汉哈萨克双语电子词典是哈萨克语自然语言处理的一个重要基础性工作，对将来的汉哈萨克机器翻译系统，哈萨克语和其他语言的双语研究都有一定的价值。

哈萨克语作为我国新疆地区的一个主要的少数民族语言，同时和新疆的维吾尔语，柯尔克孜语等都同属一个语系，彼此之间有一定的联系，在社会生活的许多方面都有着广泛的应用[3]。新疆伊犁、塔城、阿勒泰、昌吉和哈密等地有许多的哈萨克语进行教学的中小学，在进行日常教学工作时需要使用汉语和哈萨克语双语进行，同时更多的社会上的工作者也需要使用双语，而汉哈萨克双语电子词典可以有效地帮助他们处理查询和学习双语的作用。

在民族教育领域，汉哈萨克双语电子词典在新疆少数民族地区具有很好的应用前景。该电子词典可以帮助双语学习的学生，起到辅助教学的作用，并为下一步开发汉哈萨克英语的电子词典打下一定的基础。

1.2 本文的主要内容与意义

电子词典（Electronic Dictionary）近些年已经成为一个新兴的学科。它是计算机学科、语言学、语料库语言学和词典学等多学科的交叉学科，同时它也是机器翻译系统中最重要的一个环节[4]。电子词典和我们平常所见的纸质文本词典不同，电子词典的载体多为计算机等电子设备，利用电子设备自带的系统平台可进行阅读和查询等多种操作。在机器翻译系统中，电子词典的作用主要体现在它是

某种或某些自然语言的词汇信息库，对各个词汇的各项特征包括语音、形态、词义等进行详尽的描述，也可以对词汇在句子中的用法等语法信息进行描述和举例说明[5]。

电子词典在技术上的最主要的问题就是如何高效的利用有限的存储和高效查询词汇的问题。因为本文主要讨论的是文本信息的电子词典，因此文本的解压缩，加密和快速查找是重要的内容。开发电子词典，词库的整理是必不可少和需要耗费工作量的阶段，词库的建设也是需要讨论的问题[6]。

上个世纪 40 年代末美国对机器翻译的研究过程中提出了电子词典的概念，后来逐渐被人们所重视，但由于技术方面的原因，发展缓慢。20 世纪 80 年代末开始，随着个人电脑 PC 机的出现和电子技术的快速发展，以及机器翻译研究的进一步发展，电子词典的研究和应用也逐渐深入[7]。

上世纪五六十年代，关于电子词典的研究开始出现，80 年代开始逐渐得到了重视，中文信息处理方面的学者们进行了多种尝试，我国学者提出电子词典中每个记录着的静态信息都是对汉语句法和语义分析的重要基础，在中文信息处理中所有使用的运算信息都是从电子词典中的词汇所获得的。在自然语言处理系统中除了要寻找语言的句法规则外，还需要对自然语言中的词汇进行全面，深入的挖掘整理，寻找其规律[7]。同时 90 年代开始应用于自然语言处理的电子词典的研究被列入国家的中长期规划，同时进行了多方面的基础研究，例如：《信息处理用现代汉语词汇研究》、《现代汉语语法信息词典》等，其中北京大学的《现代汉语语法信息词典》对汉语句子的自动分析和生成进行了深入的研究，具有很高的质量，代表了我国在自然语言处理领域的水平。同时也开发出了许多实用性强的电子词典例如：《中国大百科全书》、《金山词霸》等，这些产品取得了极大的成功[8]。

国内的少数民族电子词典的研究领域，新疆大学、内蒙古高校和西藏大学的研究都取得了一定的成绩，例如内蒙古大学设计的《英蒙汉电子词典》，实现了英蒙汉三语的查询，词库的自动产生等。《达日罕汉蒙词典》是一部汉蒙双语词典，特点是操作简单，查询速度快。新疆大学的信息学院研究的《碧黎库特英汉维电子词典》则是实现了多种媒体的综合应用，取得了丰硕的成果[8][9]。

除了互译用的电子词典外，语法词典、词汇搭配词典等表达更多词汇特征信

息的词典也将是未来各种语言研究的热点和方向。

1.3 课题研究的关键问题

汉哈萨克双语词汇信息的录入，本电子词典的用户对象为大中专院校的学生、中小学学生和社会用户，因此我们选择的是国内最权威的《汉哈辞典》作为核心的词库，录入近 5000 条左右的词汇信息。

汉哈萨克双语词汇信息的一致性，因为是双语互译系统所以要保证汉语和哈萨克语之间的互相对应。在汉哈萨克词库建立的基础上，通过程序生成哈萨克汉语词汇库，建立一一对应关系。

实现汉哈萨克双语词汇查询功能，允许用户通过输入汉语或哈萨克语词汇，查找到对应语言的解释。实现汉哈萨克双语电子词典的维护功能，通过添加、修改和删除功能实现用户随时可以对汉哈萨克双语电子词典的维护工作。实现汉语和哈萨克语的正确显示，因为汉语是自左向右书写，而哈萨克语是自右向左书写，因此在词典显示过程中要作相应的处理。实现汉哈萨克双语电子词典的系统托盘效果，在程序后台处理时能够让软件最小化到任务栏区域，使用时通过点击图标恢复。

设计实现屏幕取词技术，通过这一技术实现即点即显得效果。在实现屏幕取词功能时，重点要考虑汉语的分词问题，在这里本文中采用的是中科院的 ICTCLAS 系统提供的接口来实现分词。哈萨克语在屏幕取词时，要实现词干提取功能，在本文中主要采用的是新疆大学提供的词干提取类来实现。这些关键问题在后面的相应章节中将详细介绍。

1.4 本论文的组织形式

全文的内容共分为五章，内容如下：

第 1 章：绪论，通过对本论文中所要研究的问题的背景和意义进行介绍，和对电子词典中相关的技术问题进行阐述，以及国内外研究现状和历史情况的介绍，得出本文的主要工作内容。

第 2 章：电子词典相关概念和技术介绍，主要阐述电子词典研究中的一些基本概念，技术和算法。

第 3 章：汉哈萨克双语电子词典系统需求分析和哈萨克语特点，介绍了汉哈萨克双语词典功能性需求和非功能性需求以及哈萨克语的特点和汉哈萨克双语电子词典中中需要用到的 Unicode 代码和词干提取等问题以及汉语的分词技术。

第 4 章：汉哈萨克双语电子词典的总体设计和相关算法介绍。

第 5 章：主要介绍了汉哈萨克双语电子词典的运行结构和功能结构等。

第 6 章：总结和展望，提出汉哈萨克双语电子词典研究中存在的不足之处和将来进一步研究的方向。

第二章　系统相关技术介绍

汉哈萨克双语电子词典在整体的开发过程中，遵循软件工程思想，同时在电子词典的设计开发过程中首先汉文处理中需要涉及到分词技术；其次哈萨克文字符因不同的输入法间采用的 Unicode 编码不尽相同，需要进行编码的统一；再次哈萨克文词汇作为典型的黏着语类型，词汇的变化都体现在附加成分中，所以需要进行词干提取处理；还有针对于电子词典本身特有的压缩算法，排序算法和查询算法的讨论及对开发语言的选择。下面将一一进行介绍。

2.1 电子词典

2.1.1　电子词典软件工程

电子词典首先应该确定所要应用的领域和对象，然后对计算机专业和自然语言处理领域中的知识结构进行组织，其次确定电子词典的整体框架结构，再将具体的词汇和词义进行录入，编撰，校对，加工，构造出词汇库的结构，对查找算法进行设计，与数据库进行关联，实现每个功能模块，对其进行单元测试，将各个模块进行整合，进行综合测试，最终通过验收测试[10]。

2.1.2　电子词典内字符编码的统一

电子词典中的词汇库的建立需要录入大量的信息，主要通过以下几种方式来收集词汇信息：第一种通过专门的排版系统软件录入的词汇信息，由于排版系统所采用的编码系统与现有的 Unicode 编码有所差异，所以在从排版系统获得的词汇库需要经过一次代码转换的问题；第二种是通过扫描文本词典，然后利用光学字符识别（OCR）系统来转化，随后进行校对编辑；第三种是通过网络获得的词汇信息[8]。随着信息化的发展和网络的日益普及，使用标准化的字符编码是大势所趋，因此在本系统中主要采用的是 Unicode 编码[11][12][13]。

2.1.3　电子词典词汇库

词汇库对于电子词典的作用不言而喻，拥有丰富的词汇信息对于词汇库建设